Análisis Taller 3

Sebastián Henao Pinzón, David Derby Cardona
Grupo 5, Laboratorio 2
Universidad de los Andes, Bogotá, Colombia
{s.henao41, d.derby10}@uniandes.edu.co

Fecha de presentación: Abril 10 de 2014
Fecha de presentación del bono: Abril 21 de 2014

Forma de acceso a la aplicación Web

La aplicación web está disponible en http://10.0.1.196/.
Se puede seleccionar una consulta para ejecutar o seleccionar el enlace de la página de visualización.

Resultados del procesamiento

Tras todo el procesamiento sobre el dataset, se obtiene:

- 3237 tweets procesados
- 1610 tweets georeferenciados
- 61.81% de PRECISION al hacer el análisis de sentimiento.

Descripción de la solución

Los retos del taller se abordaron en dos fases: Carga y Consumo.

Durante la fase de carga, se buscó preprocesar los tweets para facilitar el consumo web posterior y enriquecerlos con datos de geolocalización. En particular después de resolver el proceso de leída de tweets del dataset, para cada tweet:

- Se le generó un 'summary rating' que intentaba reconciliar las posibles diferencias de opinión entre los evaluadores.
- Se le calculó un valor de 1 a 5 de sentimiento (se usó el modelo de clasificación para texto genérico en inglés de la Universidad de Stanford[1])
- Usando una versión reducida del dataset de ciudades de MapMind[2] y con ayuda de un Trie con los nombres de ciudades estadounidenses y países, se encuentran posibles candidatos para geolocalización.

1http://www-nlp.stanford.edu/software/corenlp.shtml

2http://www.maxmind.com/en/worldcities

- Se extraen los hashtags con una expresión regular simple (“ #(\w*) “)

El ‘summary rating’ se usó para calcular un valor de PRECISION del modelo clasificador. El modelo usado alcanzó una precisión del 61.81%, seis puntos por debajo del estándar de industria.

El Trie permitió usar listas grandes (la lista de lugares es de más de 50000 sitios geolocalizados) para hacer matching palabra por palabra del tweet sin poner más carga que la que es consumida por el proceso de análisis de sentimiento.

Para la fase de consumo, se decidió aprovechar la alternativa que Mongo ofrece a MapReduce: Consultas agregadas. Dado que los requerimientos eran variaciones de conteo (hashcloud, geocloud), consultas como:

```
db.tweet_main.aggregate([
   { "$project": { "loc":1 }},
   { "$unwind": "$loc" },
   { "$group": { "_id": "$loc", "count": { "$sum": 1 } }}
])
```

Permiten ser “complejas” (para cumplir uno de los requerimientos del taller) y resolver las preguntas que hacemos.

La aplicación web fue desarrollado en PHP. El mapa que visualiza los tweets geocodificados con el API de MapQuest[3] que utiliza los mapas de OpenStreetMap. MapQuest no funciona con tantos sitios/ubicaciones entonces fue necesario limitarlo a 50 sitios para que el mapa muestre algo.

Análisis de Resultados

Para la fase de carga:

- En la Máquina Virtual (4 procesadores, 2GB RAM), el proceso tarda 5:41.72 en terminar.
- En el clúster, el proceso tarda 5:32 en terminar

El peso principal del proceso de carga es el análisis de sentimiento, siendo la “caja negra” del proyecto.

Bono

Para tomar una muestra de tweets de Colombianos utilizamos la librería Twitter4j[4]. El dataset nuevo fue guardado en un archivo *tsv* en el mismo formato que tiene el dataset original para que la aplicación pueda procesarlo sin muchos cambios. El dataset fue generado por once tuiteros Colombianos diferentes y resultó con 436 tweets durante los últimos 10 días. Había problemas de correr la aplicación con Mongo en e clúster y por

3 http://www.mapquestapi.com/staticmap/wizard.html

4 http://twitter4j.org

eso no pudimos hacer el análisis.